강우원, 김다윤, 김소연, 정용헌

NLP 감정분석 기반
마케팅 시장 분석

NLP 감성 분석에 기반한 온라인 커뮤니티 이슈 키워드 모니터링 시스템

빅데이터미네이터

“ ☹😐☺

커뮤니티별 주요 이슈 키워드 감성 분석 시스템 ”

- 선정한 4가지 온라인 커뮤니티 인기게시판에서 제목, 본문, 댓글을 크롤링하여, 1~10위 이슈 키워드 추출
- 10가지 이슈 키워드가 포함된 제목, 본문, 댓글만 다시 수집하여 전처리 진행
- 전처리 된 데이터에 대한 감성 분석 진행 후, 그 결과를 시각화하여 제공

⇒ 커뮤니티 별 이슈 키워드와 이에 대한 감성 분석 결과를 제공하는 마케팅 인사이트 툴

온라인 커뮤니티 - 4가지 선정

실시간 커뮤니티 순위 (2021-06-10 ~ 2021-06-16)

순위	커뮤니티	투베스 지표	글수	조회수	반응수	댓글수
1	FM	81.661	438	73,921,430	449,702	133,766
2		68.312	2,755	117,108,435	364,304	188,994
3	NATE 판	45.806	1,123	46,903,697	275,669	110,850
4	dcinside.com	29.096	892	36,454,992	455,508	315,987
5	RULIWEB	22.937	816	31,030,273	116,194	45,526
6		18.714	531	20,361,983	0	305,876
7	CLIEN.net	15.646	1,824	21,690,383	114,393	47,613

- 인기 온라인 커뮤니티 10위 내에 있는 4가지 커뮤니티 선정
  → 충분히 크롤링 할 수 있을 만큼 많은 사용자들이 게시물을 올려야 함

- 다양한 사회적 이슈를 모아서 보여주는 베스트 게시판 존재
  → 의미 있는 이슈 키워드와 감성 분석 결과를 도출하기 위해, 다양한 사회적 이슈를 다루는 인기 게시물이 모여 있어야 함

- 다양한 의견을 보여주기 위해, 정치 성향과 성별을 고려하여 커뮤니티 선정

https://todaybeststory.com/ranking_current.html

커뮤니티 - 게시판 선정

- 네이트판 : 톡커들의 선택 명예의 전당
  → 조회수, 추천수, 댓글 등을 고려한 베스트 게시글이 모여 있음

- 루리웹 : 정치유머 베스트 게시판
  → 가장 추천을 많이 받은 정치, 유머 게시물이 모여 있음

- 에펨코리아 : 포텐 터짐 게시판 화제순
  → 조회수, 추천수, 댓글 등을 고려한 가장 화제가 되는 게시물이 모여 있음

- 클리앙 : 공감글 게시판
  → 가장 공감을 많이 받은 게시물이 모여 있음

전단부 구조

1. Input

- 커뮤니티

→ 네이트판, 루리웹, 에펨코리아, 클리앙

- 크롤링

→jupyter notebook 환경에서 Beautiful soup 라이브러리를 이용하여, 각 커뮤니티 사이트의 html, css 요소 등에 접근하는 방식

- 제목, 본문, 댓글 크롤링하여, 전날 이슈 키워드 추출

2. 전처리

- 영어, 특수문자, 초성 등 불용어 제거

- 형태소 단위로 토큰화

→ Korean NLP인 KoNLPy, Komoran, Okt 형태소 분석기 활용

→ 한국어 자동화 표준 띄어쓰기를 위해 Pykospacing 활용

- csv 파일로 저장 (프로토타입, 추후 DB 구축)

키워드 추출

키워드 추출

“TF-IDF 텍스트 마이닝”

1. 문서 내 단어들의 척도를 계산하여 핵심어 추출
2. 특정 단어가 문서 내에서 얼마나 중요한지에 대한 척도 계산

TF : 특정 문서 d에서의 특정 단어 t의 등장 횟수
DF : 특정 단어 t가 등장한 문서의 수
IDF : DF값의 역수(반비례)
TF-IDF : TF와 IDF를 곱해준 값

$$tf\,idf(t,d,D) = tf(t,d) \times idf(t,D)$$

$$tf(t,d) = \log\left(f(t,d) + 1\right)$$

$$idf(t,D) = \log \frac{|D|}{|\{d \in D : t \in d\}| + 1}$$

⇒ 특정 문서 내에 단어 빈도가 높을 수록, 그리고 전체 문서들 중 그 단어를 포함한 문서가 적을수록 TF-IDF 값이 높게 나온다.
즉, 해당 단어가 그 특정 문서에서 키워드로 작용을 한다는 의미이다.

키워드 추출

1. 이종화, 이문봉, 김종원. (2019). TF-IDF를 활용한 한글 자연어 처리 연구. 정보시스템연구, 28(3), 105-121.

→ TF 값만으로 한 문장 내의 단어와의 연관성을 나타내기 힘들며, 오류 발생 가능
→ 이러한 TF값의 치명적 오류를 바로 잡기 위하여 IDF 활용
→ 연관성 없는 단어들에 제한을 주기 위한 기법으로 TF-IDF 활용

사 'a'처럼 연관성 없는 단어가 발생된다. 즉, TF 값만으로 한 문장내의 단어와의 연관성을 나타내기 힘든 결과를 얻게 된다. 단순 단어 빈도가 높다고 문장의 연관성을 높게만 판단하기엔 오류가 발생할 수 있다. 특정 단어가 문서나 문장의 전체에서 얼마나 공통적으로 나타나는지를 확인하여 문장 내 자주 등장하는 단어를 연관성 없는 단어들에 제한할 필요가 있다.

2. 곽수정, 김현희. (2019). 텍스트 마이닝과 토픽 모델링을 기반으로 한 트위터에 나타난 사회적 이슈의 키워드 및 주제 분석. , 8(1), 13-18.

→ TF-IDF 기법을 사용하여, 커뮤니티와 비슷한 성격인 SNS에서 사회적 이슈 키워드 추출
→ TF-IDF는 중요한 단어에 가중치를 부여하여, 문서 내에서 해당 단어가 얼마나 많은 비중을 차지 하는지 알 수 있기 때문에 보다 정확한 중요도 파악 가능

를 시행하였다. TF-IDF는 TF(Term Frequency)는 특정 문서 하나에서 특정 단어가 나온 횟수를 나타내고, IDF(Inverse Document Frequency)는 특정 단어의 전체 문서내의 빈도를 역수로 취한 값이다[8]. 즉, TF-IDF는 단순 빈도에 가중치를 부여하여 문서 내에 얼마나 많은 비중을 차지하는지 나타내기 때문에 보다 정확한 중요도를 파악할 수 있다.

키워드 추출

“ TF-IDF ”

용어 가중치 기법이 검색 성능을 향상
타 모델에 비해 더 많은 양의 정보 내 문서 비교가 가능
단순하고 빠르기 때문에, 현재 가장 대중적인 검색 모델

선정한 커뮤니티 4곳에서 얻은 여러 게시물 사이의 유사도 및 중요도를 비교하기 적합하다.
또한 데이터 양이 많기 때문에 TF-IDF가 속도 면에서도 빠르고 적합한 추출 기법이다.

키워드 추출

1) 선정한 커뮤니티의 인기 게시판에서 제목, 본문, 댓글을 크롤링 한 후, 전처리를 거쳐 명사화 함

2) TF-IDF 기법을 적용하여, 각 커뮤니티 별 탑 10 키워드 추출

```python
#TF-IDF 이미 한번 명사 단위로 정리 한 명단으로 처리할 시

as_one = ''
for noun in nouns:
    as_one = as_one + ' ' + noun
words = as_one.split()

counts = Counter(words)

vocab = sorted(counts, key=counts.get, reverse=True)

#단어 빈도 정리
word2idx = {word.encode("utf8").decode("utf8"): ii for ii, word in enumerate(vocab,1)}

#딕셔너리로 정리
idx2word = {ii: word for ii, word in enumerate(vocab)}
```

```python
#tf-idf
tfidf = TfidfVectorizer(max_features = 10, max_df=0.95, min_df=0)

#generate tf-idf term-document matrix
A_tfidf_sp = tfidf.fit_transform(nouns)  #size D x V
```

```python
#tf-idf dictionary
tfidf_dict = tfidf.get_feature_names()
print(tfidf_dict)
```

```
['간부', '군대', '기부', '나라', '백신', '병원', '수술', '아웃', '추신수', '후원']
```

커뮤니티 별 Top 10 키워드 추출

데이터 추출 일자 : 2021. 05. 22 – 2021.06.06

1. 네이트판

['결혼', '남자', '남편', '본인', '부모', '시간', '엄마', '여자', '자기', '친구']

2. FM 코리아

['간부', '군대', '기부', '나라', '백신', '병원', '수술', '아웃', '추신수', '후원']

3. 루리웹

['매수', '민주당', '부모', '사건', '새끼', '유튜브', '의대', '자식', '친구', '한강']

4. 클리앙

['국가', '국민', '나라', '대만', '대통령', '백신', '일본', '중국', '처리', '한국']

감성 분석

KNU 감성어 사전

- 오픈 한글, SentiWordNet 등 다양한 범용 감성어 사전, 서비스 종료 혹은 번역으로 인해 한국 감성 어휘의 특징 반영되지 않음
- 군산대학교 소프트웨어융합공학과 Data Intelligence Lab에서 개발한 기초자료로, 한국어 범용 감성어 사전

- **KNU 감성어 사전 선정 이유:**

1. 한국어에 적합한 감성어 사전

   표준국어대사전의 뜻풀이의 감성을 Bi-LSTM을 활용하여 89.45%의 정확도로 분류하는 군산대의 한국어 감성어 사전

2. 감성, 어구, 문형 등 다양한 형태의 적용

   긍정, 부정, 중립에 대한 감성 어휘를 1-gram, 2-gram, 어구, 문형 등 다양한 형태로 추출 가능

3. 온라인 감성 어휘의 포함

   온라인 텍스트 데이터에서 사용되는 신조어, 이모티콘에 대한 감성 어휘도 포함

4. 도메인의 제약이 없다

   도메인에 영향을 받지 않는 사전으로 타 사전에 비해 정확도가 높다

5. 새로운 감성어 추가에 적합

   리커트 척도를 이용하며, 개발자 (평가자)들의 합의를 통해 각 단어의 긍부정이 판별된 사전

   => 따라서 해당 프로젝트에서 새로운 단어를 감성어 사전에 추가 할 경우에도 이러한 척도를 사용 할 수 있다.

KNU 감성어 사전 활용 방법

1. KNU 사전은 단어의 입력을 통해 해당 단어의 형태 혹은 극성 (감성) 정도 값을 출력한다.

2. 해당 프로젝트에서는 크롤링한 데이터를 전처리 및 토큰화 한 이후,
제공되는 SentiWord_Dict를 활용하여 긍, 부정 감성 점수를 합산한다.

3. 온라인 커뮤니티에 대한 감성 분석이므로 해당 사전에
새로운 온라인 용어나 이모티콘 등의 감성어를 추가하여 사용한다.

<u>새로운 감성어 추가 기준</u>

- (1) 매우 부정 (2) 부정 (3) 중립 (4) 긍정 (5) 매우 긍정 등 리커트 척도를 사용한다.

- 프로젝트 진행 팀원 4명이 평가자가 되어 각 단어의 긍정, 중립, 부정을 판별하고,

이의가 있을 경우 표준어 대사전 등을 참고하며, 토론을 통해 합의를 이루는 방식 사용

(voting 방식)

- 긍부정어 통계

긍부정어	단어개수
1-gram 긍부정어	6,223
2-gram 긍부정어	7,861
긍부정 어구	278
긍부정 문형	253
긍부정 축약어	174
긍부정 이모티콘	54
1-gram 긍부정어	6,451
2-gram 긍부정어	8,135
3-gram 긍부정어	226
4-gram 긍부정어	20
5-gram 긍부정어	5
6-gram 긍부정어	3
7-gram 긍부정어	2
8-gram 긍부정어	1
매우 긍정(2)	2,597
긍 정(1)	2,266
중 립(0)	154
부 정(-1)	5,029
매우 부정(-2)	4,797

SentiWord_Dict.txt

```
가늘 수 없게 -2
가늘 수 없을 -2
가늘 수 없이 -2
가능성이 늘어나다 2
가능성이 있다고 2
가능하다 2
가닐가닐하다 -1
가다듬어 1
가다듬어 수습하는 1
가다듬어 수습하다 1
가다듬어 정하다 1
가당찮다 -2
가당찮이 -2
가당히 1
가두거나 -1
가두거나 하치거나 -2
가드름려 -1
가뜬하게 1
가라앉다 0
가라앉지 않은 0
가라앉혀 바로잡다 1
가래 -1
가래 따위가 -1
가래가 -1
가래가 섞인 -1
가량맞고 -1
가량맞고 야살스러운 -1
가량맞고 야살스러울 -1
가려서 좋아하다 -1
가려운 -1
가려운 느낌이 -1
가려운 증상을 -1
가련하게 -1
가련하게 여기다 -1
가련하게 여김 -1
가련한 -1
```

프로토타입 감성 분석

1. 크롤링을 통해 각 커뮤니티 별로 데이터를 축적해 두고,

전체적으로 여러 키워드 추출 및 감성 분석 기법을 시도하여 개발

- 데이터 추출 일자 : 2021. 05. 22 – 2021.06.06

2. **각 커뮤니티 별**로 10개의 키워드를 추출하고,

이에 대한 감성 분석 결과를 도출

- 전날 데이터를 정리하여, 다음날 사용자에게 제공하고자 하므로,

하루 치 데이터를 추출하고 분석하는 방식으로 개발

전체 구조

키워드 추출 ~ 감성 분석 - 예시

1. 선정 게시판 크롤링

fm.csv

2. 키워드 Top 10 추출

['간부', '군대', '기부', '나라', '백신', '병원', '수술', '아웃', '추신수', '후원']

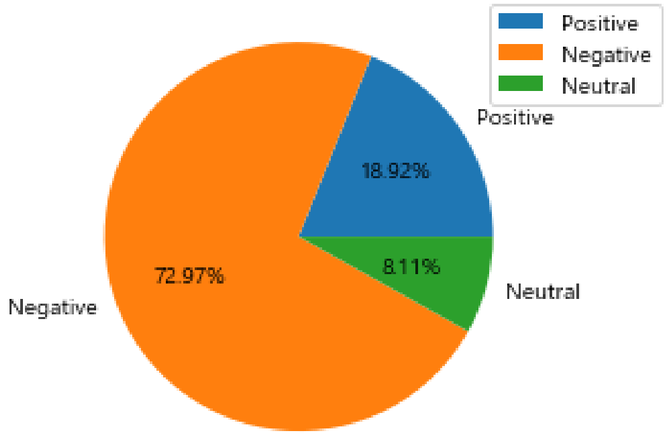
KNU 감성어 사전 활용하여 감성 분석

3. 각 키워드에 대한 데이터만 추출 및 토큰화

	index	제목	본문	댓글
0	5	오늘자 콜센터 상담사의 외 침	한남들 후원 취소하는 거 타격 도 없으니까 홈페이지로 취소 문의 넣어 달래	저런 거 보면 그냥 장작 넣을려고 위장 같음 드가 자재는 초록우산 측에서 백퍼 고소...
1	7	이 기적인 한남들 기부해봐야 얼마나 한다고	요즘 핫한 초록우산 페미 지원 뉴스베댓 한남들에게 일 침 남자 놈들 기부 해봐야 얼...	솔직히 말해서 여초애들이 후원을 하겠냐고 족보세트 하나라 더 시켜 먹으려고 할 건데...
2	9	초록우산 문의 답변 받음	내가 초록우산 한 년 넘게 후원하고 있는데 어제 초록우산 관련 글 보고 바로 문의 ...	그 디씨에 올라온 상담사 글이 일 수도 있겠네 한 남꺼져 이거구만 후원자가 하는 질...

4. 감성 분석

	Keyword	negative	neutral	positive
0	백신	-27	3	7

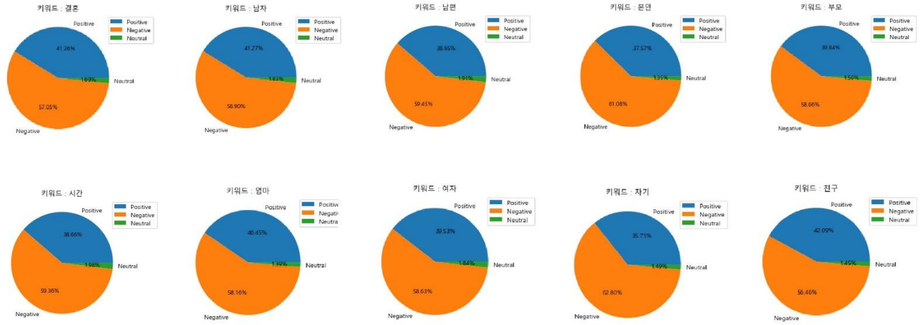


커뮤니티 별 감성 분석 결과

톡커들의 선택 명예의 전당

네이트판

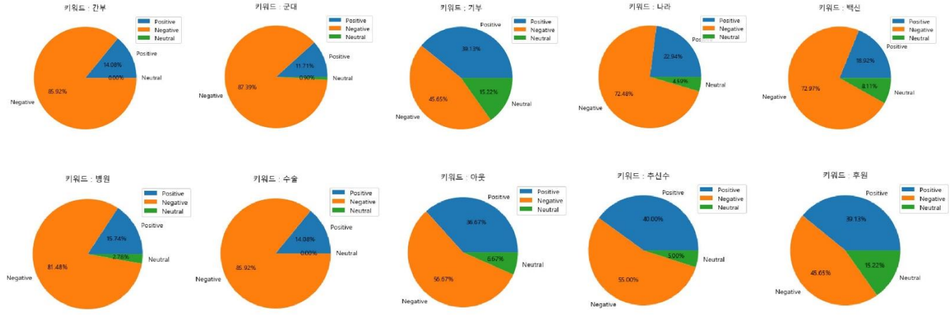
키워드: ['결혼', '남자', '남편', '본인', '부모', '시간', '엄마', '여자', '자기', '친구']



포텐 터짐 화제순

FM 코리아

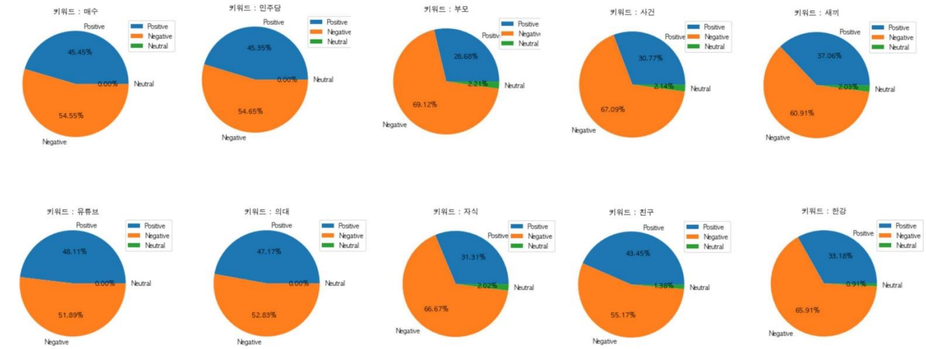
키워드: ['간부', '군대', '기부', '나라', '백신', '병원', '수술', '아웃', '추신수', '후원']



정치유머 게시판

루리웹

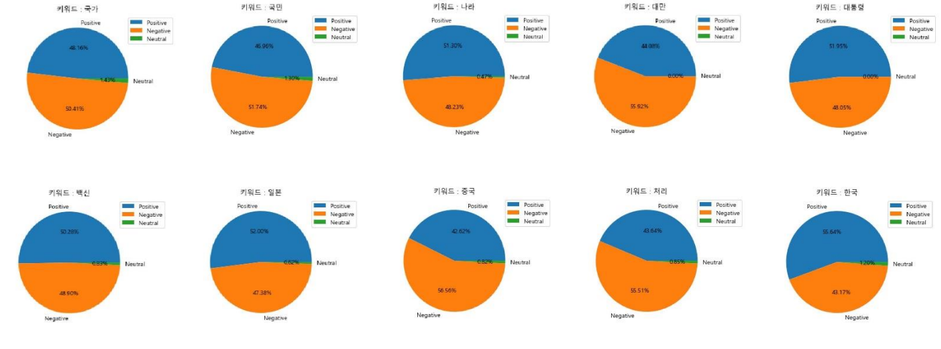
키워드: ['매수', '민주당', '부모', '사건', '새끼', '유튜브', '의대', '자식', '친구', '한강']



CLIEN 공감글

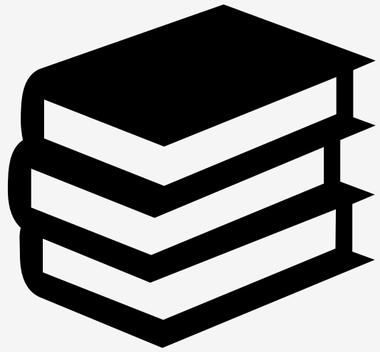
클리앙

키워드: ['국가', '국민', '나라', '대만', '대통령', '백신', '일본', '중국', '처리', '한국']



향후 계획

감성어 사전 추가 구축



SentiWord_Dict.txt

```
가볍게 행동하거나	-1
가볍게 행동하는	-1
가볍게 행동하다	-1
가볍고	1
가볍고 보드랍게	1
가볍고 상쾌하다	2
가볍고 상쾌한	2
가볍고 시원하게	2
가볍고 편안하게	2
가볍고 환하게	2
가분가분	1
가분히	1
가뿐가뿐	1
가뿐가뿐하다	1
가뿐가뿐히	1
가뿐하게	1
가뿐하다	1
가뿐한	1
가뿐한 느낌	1
가뿐한 느낌이	1
가뿐히	1
가쁘게	-1
가쁘게 쉬다	-1
가쁘고	-1
가쁘고 거칠게	-1
가쁘고 급하게	-1
가쁜	-1
가쁜 증상	-1
가살스럽다	-1
가소롭게	-1
```

KNU 한국어 감성어 사전에 없는 인터넷 용어나 난독화 된 단어들을 정리하고,

앞서 정한 Voting 기준에 따라 감성어를 추가하여,

분석의 품질, 즉 **정확도**와 **분석력**을 높이는 것을 목표로 하여 개발을 진행할 것이다.

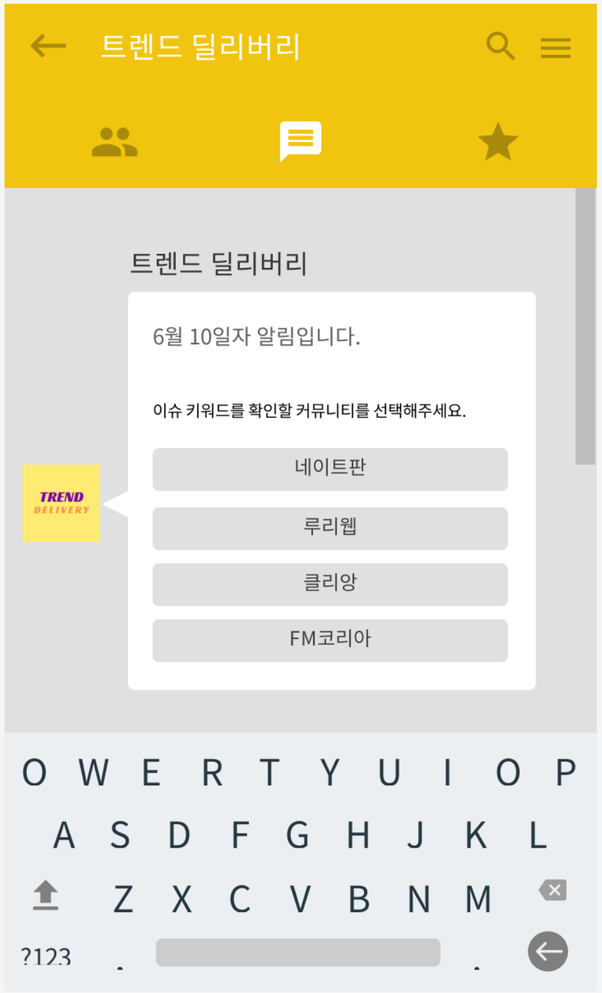
키워드 추출 방식 보완

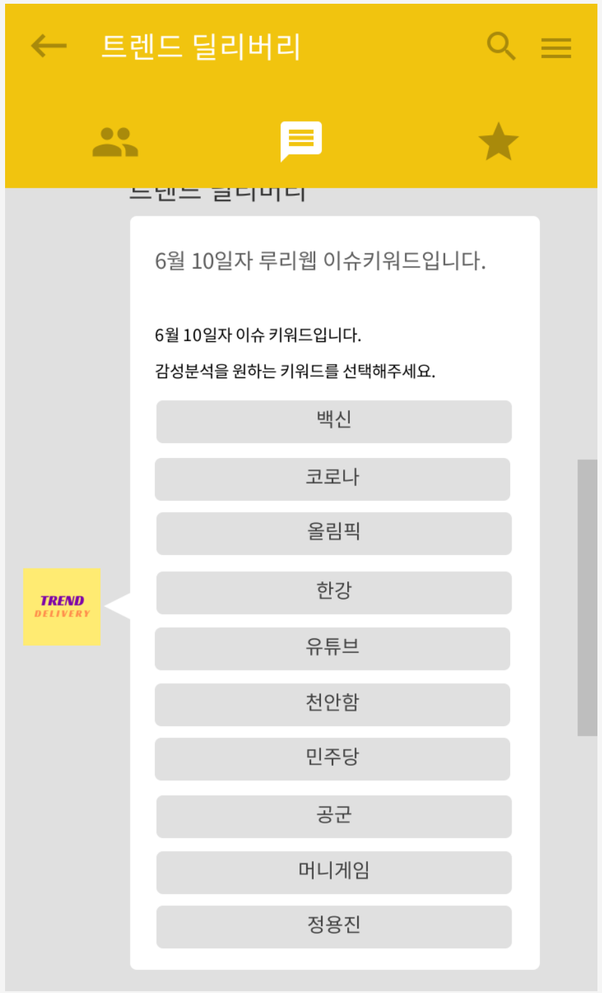
현재 프로토타입에서 사용 중인 TF-IDF 방식은 유사어 분류에 약하다는 단점이 있다.
이를 보완하기 위하여, 자주 혼용되는 아래의 기법들을 추가적으로 실습해보고 결과를 비교하여 사용할 예정이다. 이 과정을 통해 더 정확하고 유의미한 키워드를 추출하고자 한다.

- 토픽모델링의 일종인 LSA (Latent Semantic Analysis)
: 동음이의어 등 유사도 측정하여 의미론적으로 같은 내용을 묶어주는 방식

- Word2Vec
: 벡터 기법 중 하나인 일종의 얕은 신경망으로, 단어들 간의 유사성을 표현하는 방식

최종 목표: 챗봇 서비스





현재까지 제작된 프로토타입 모듈을 기초로 하여, 하나의 프로그램으로 웹페이지를 개발하고 관리자 시스템을 생성할 것이다. 사용자에게 챗봇 형식으로 매일 주기적인 시간에 알림을 제공하여, 전날의 이슈 키워드와 감성 분석 결과를 확인할 수 있게 해 주는 서비스를 제공할 계획이다.

개발 환경

개발 언어: python,

Back-End 프레임워크: django (장고, 파이썬 기반)

Front-End 프레임워크: BOOTSTRAP

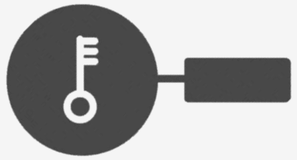
DB는 django 프레임워크에 기본적으로 지원되는 SQLite3를 사용합니다.

- 부트스트랩(Bootstrap)은 웹사이트를 쉽게 만들 수 있게 도와주는 프론트엔트 프레임워크로, 하나의 CSS로 휴대폰, 태블릿, 데스크탑까지 다양한 기기에서 작동하며 사용자가 쉽게 웹사이트를 제작, 유지, 보수할 수 있도록 기능을 지원하기 때문에 선택하였다.

- django는 파이썬으로 작성된 오픈 소스 웹 애플리케이션 프레임워크로, 쉽고 빠르게 웹사이트를 개발할 수 있는 구성요소로 이루어져 있다. 또한 파이썬 언어를 기반으로 하여 다양한 라이브러리들을 그대로 사용할 수 있다는 장점을 가지고 있기 때문에 채택하였다.

기대 효과

정치, 사회, 문화 등 다양한 분야에서의 온라인 여론 분석을 통해
체계적인 마케팅 전략을 수립할 정보 생성



01

매일 온라인 커뮤니티
이슈/여론을
간편하게 파악 가능

02

빅데이터 분석을 기반으로
추출한 키워드 및 감성 분석
데이터에서 인사이트 도출

03

도출한 인사이트를 바탕으로
맞춤형 마케팅 전략 수립 가능

감사합니다 :)

빅데이터미네이터

손윤식 교수님 주식회사 닥터송 송환구 멘토님